DOCUMENT RESUME

ED 219 417 TM 820 428

AUTHOR Kuntz, Patricia
TITLE Test-Wiseness Cues in the Options of Mathematics Items.
PUB DATE Mar 82
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Cues; *Item Analysis; *Multiple Choice Tests; Psychometrics; *Quantitative Tests; Statistical Analysis; Test Coaching; *Test Construction; Testing Problems; Test Items; Test Validity; *Test Wiseness

ABSTRACT

The quality of mathematics multiple choice items and their susceptibility to test wiseness were examined. Test wiseness was defined as "a subject's capacity to utilize the characteristics and formats of the test and/or test taking situation to receive a high score." The study used results of the Graduate Record Examinations Aptitude Test (GRE) and Graduate Management Admission Test (GMAT). Principles of test wiseness, specifically for mathematics items, were developed and applied. A large proportion of the items studied were susceptible to test wiseness at a significantly higher level than expected by chance. The study investigated the possibility that students could be trained to detect flaws in test items and profit from them. The four and five option items were examined with respect to their susceptibility to test wiseness and their corresponding psychometric properties. Two methods of reducing test wiseness were training all potential examinees in test wiseness or elimination of susceptible items. The reduction of options from five to four would seem to reduce the susceptibility of items to test wiseness without affecting the psychometric properties. (DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED219417

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Test-Wiseness Cues in the Options of Mathematics Items

Patricia Kuntz

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. Kuntz

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A Paper Presented at the 1982

American Educational Research Association Annual Meeting

TM 820 428

Abstract

Literature on test-wiseness has shown that verbal items are susceptible to test-wiseness; this paper is interested in whether mathematics items are susceptible to test-wiseness and if test-wiseness cues can be eliminated by reducing the number of options. Principles of test-wiseness specifically for mathematics items were developed and applied to several math tests. Five of six tests studied were susceptible to test-wiseness at a level significantly higher than expected by chance. It was also concluded that these principles are trainable and that test-wiseness cues can be eliminated without affecting various psychometric properties of a test by changing 5-option items to 4-option items.

Introduction

Robert Ebel writes that multiple-choice test items "are adaptable to the measurement of most important educational outcomes--knowledge, understanding, and judgement; ability to solve problems, to recommend appropriate action, to make predictions" (1965, p. 149). We have seen multiple-choice tests come into widespread use during the last few decades because "there tends to be less indirectness and artifice in multiple-choice than some other item forms" (Ebel, 1965, p. 150).

This study examines the quality of multiple-choice items on some well known and frequently used mathematics tests. Specifically, this study is interested in questions concerning the susceptibility of mathematics items to test-wiseness: Are mathematics items flawed in that the options reveal the correct answer to the test-wise examinee? If these items are flawed, can the naive examinee be trained to select the correct answer based on properties of the options? Would reducing the number of options from five to four reduce the susceptibility of these tests to the identified flaws? For the purposes of this study, test-wiseness will be defined as "a subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score" (Millman, Bishop, and Ebel, 1965, p. 707).

Jeffrey K. Smith best phrases the initial conceptualization of this study: "I began my research on this topic when I discovered a peculiar but pervasive characteristic of multiple choice items: many multiple choice items can be answered by simply looking at the alternatives and not reading the question at all!" (1980, p. 4). Upon writing

and reviewing mathematics items, the experimenter began to recognize some general patterns formed by the numbers in the options of these items. Often the correct answer could be isolated before the stem of the item was even read. Also, when writing items, the experimenter often experienced difficulty in constructing four distractors, and found that often options in the items she wrote formed patterns indicative of the correct answer. This led to the questions: Are mathematics items susceptible to test-wiseness? Can examinees also recognize patterns in the options and select the correct answer from just looking at these patterns?

In addition, the experimenter wondered whether reducing the number of options from five to four would reduce the susceptibility to test-wiseness. She found that with an odd number of options, there was often an imbalance among the options, which pointed to the correct answer.

The questions raised above suggested the study reported here. The investigator sought additional data on the prevalence of mathematics items with flaws leading to test-wiseness, conducted an experiment to learn if students could be trained to detect such flaws and profit from them, and compared 4- and 5-option items with respect to their susceptibility to test-wiseness and their corresponding psychometric properties. In summary, the hypotheses that this work is testing are:

(1) The options of well known and frequently used multiple-choice mathematics tests contain patterns that permit the correct answer to be identified at greater than chance levels without knowledge of the item stem.

(2) Tests with 5-option items will be more susceptible

to test-wiseness than tests with 4-option items.

(3) Without training in test-wiseness (defined here as recognizing patterns among options), subjects are expected to be able to guess the correct answer from looking at only the options 20 percent of the time for 5-option items and 25 percent of the time for 4-option items.

(4) Immediately after training in test-wiseness, subjects are expected to be able to guess the correct answer from looking at only the options significantly more than 20 percent of the time for 5-option items.

(5) Immediately after training in test-wiseness, subjects are not expected to be able to guess the correct answer from looking at only the options significantly more than 25 percent of the time for 4-option items that are derived in an attempt to eliminate test-wiseness cues.

(6) Subjects trained in test-wiseness are expected to be able to guess the correct answer from looking at only the options significantly more than the subjects not trained in test-wiseness on 5-option items, but not on 4-option items that are derived in an attempt to eliminate test-wiseness cues.

(7) It is expected that the quality of a test (with respect to an index of discrimination and reliability) with 5-option, cue-infested items will be comparable with that of a test with 4-option, cue-reduced items.

Determining the Prevalence of Mathematics Test Items That are Susceptible to Test-Wiseness

The literature on testing includes criteria for writing multiple-choice items and principles of test-wiseness (see Mosier, Myers, and Price, 1966; Millman, Bishop, and Ebel, 1965). For the most part, these guidelines are geared for verbal items. They are concerned with aspects of the options such as grammatical consistency and option length and complexity, and these guidelines are not readily adaptable to mathematics items.

To measure the prevalence of mathematics items that are susceptible to test-wiseness, it was first necessary to define those patterns among item options that will be counted as flawed. A list of the principles of test-wiseness that were developed for this study appear in Figure 1.

Figure 1

Principles of Test-Wiseness
for Mathematics Items

1. Eliminate "odd" options.
2. Select the middle value of a regular sequence of values.
3. Select the middle value of a subgroup of items with common characteristics.
4. Select options that are multiples of other options in the item.
5. If a fraction and its reciprocal both appear, select one of these.
6. If two fractions sum to 1, select one of these.
7. If two angles sum to 90°, 180°, or 360°, select one of these.
8. If two angles differ by a multiple of 360°, select one of these.
9. When variables are involved, select the option that has the most number of similarities to the other options.
10. In general, the more groups an option is a part of, the more likely it is to be the correct answer.

The development of these principles was based on observations made by the investigator during the review of hundreds of mathematics items. Several types of patterns seemed to emerge, and these were classified into the categories in Figure 1. Although the initial

conception of these principles was purely subjective, the final versions of these principles are consistent with a number of documented test-wiseness principles. A brief discussion of each principle follows.

Eliminating an "odd" option is a logical rule to follow. Millman and Pauk (1960) note that the correct option will not be a silly idea, a completely unreasonable option. In mathematics, a number that is excessively larger or smaller than the other options or in a very different form from the other options is equivalent to a silly or unreasonable statement.

The principles to choose middle values were based on experience in writing items. There seems to be a natural tendency to want to balance the options around the correct option. This is also recognized by Millman and Pauk (1969), who note that generally but not always the correct answer will not be at the extremes of options in a natural order.

The idea of looking for options that are multiples of each other parallels the item-writing caution of options similarity (Miller, Williams, and Haladyna, 1978). It is easy to see how options that are multiples of each other could be generated. Often distractors are derived by making common mistakes in solving the problem (e.g., forgetting to multiply by 2).

Similarly, the principles involving fractions and angles follow from the construction of distractors based on expected student errors. One technique for constructing distractors involves writing an option that is the correct answer to another related question (Miller, et al., 1978). Problems which involve fractions and angles often involve ratios,

proportions, measures of angles of triangles, and so on. Related questions may ask for an inverse ratio or complementary angle; if an examinee is not paying attention, these answers would appear plausible.

The final principles listed in Figure 1 follow directly from Smith's (1980) convergence strategy. Smith gives the following example of the distractors from a susceptible mathematics item taken from the Scholastic Aptitude Test:

A. $p + q$

B. $\frac{p + q}{2}$

C. $2p + 3q$

D. $\frac{2p + 3q}{5}$

E. $\frac{3p + 2q}{5}$

The following reasoning is used to select the correct answer. The majority of the options are fractions, so the correct answer is likely to be one of these (B, D, E). Two of these options contain 5 in the denominator (D, E), so these are more likely. Choosing between D and E, D would be the better choice since both C and D contain $2p + 3q$. Here, the correct answer has <u>all</u> of the most prevalent characteristics. In other words, the dimensions of the options converge on the correct answer. Smith (1980) has called this the "convergence strategy."

In order to determine whether a problem with respect to test-wiseness in mathematics items indeed exists, a number of standardized tests were analyzed. For each item, the probability of selecting each option was determined. This was done by applying as many of the principles as

possible to each item to generate the number of times each option would be selected, based on the set of principles. These frequencies were converted to percentages, which served as the probabilities. For example, in a problem with options:

(A) 6 (B) 7 (C) 9 (D) 10 (E) 12,

option (A) is a possible choice because (A) is one-half of option (E); likewise, option (E) is a possible choice. Another possible choice is option (C), since it is in the middle of the sequence 6, 9, 12. Hence, probabilities would be assigned to each of the options as follows:

(A) 0.33

(B) 0

(C) 0.33

(D) 0

(E) 0.33

For each test examined, the probabilities assigned to the correct answer of each item were summed, and this total was divided by the number of items in the test. The resulting figure was the susceptibility index of the test to test-wiseness.

Since the initial conception of the study was based on working with the Graduate Record Examinations Aptitude Test and the Graduate Management Admission Test, three forms of the GRE and two forms of the GMAT were analyzed. Only items with numerical options were examined. Items that required examinees to refer to graphs were omitted, since often these items are interrelated and contain non-numerical options. Susceptibility indices for these tests (subsets of items) were first calculated by the

experimenter and then by consensus with an independent consultant.

The principles of test-wiseness developed for mathematics items were also applied to the subtests of four other tests. These tests are the Iowa Test of Basic Skills (ITBS) (Form 2, Grades 8-9), the California Achievement Test (CAT) (Form A, Level 5--grades 9-12), the Differential Aptitude Test (DAT) (Form T), the Stanford Diagnostic Mathematics Test (SDMT) (Form A, Blue Level--grades 8-12). Only items with numerical options were examined. As a check on the experimenter's objectivity, a consultant independently applied the principles to Test 1 of the Stanford Diagnostic Mathematics Test.

Results and Discussion

Table 1 contains the susceptibility indices for the GRE and GMAT tests and indicates whether a t-test comparing the SI to the chance level is significant. The items on these two tests are 5-option items. Therefore, the expected score, by chance, is 20 percent. From Table 1, four of five tests had susceptibility indices appreciably greater than 20%. In fact, one test had an index more than twice that expected by chance. The average index for the items taken as a whole was .295. When the index for GMAT - II is excluded, this average increases to .33. Although an index of .33 does not appear to be exceedingly high, it is the result expected for a test with three-choice items and it is significant. Hence, it was judged that the susceptibility of mathematics items to test-wiseness was prevalent enough to support Hypothesis 1.

Table 2 contains the susceptibility indices for the subtests of the additional four tests to which the principles of test-wiseness for

Table 1

Susceptibility Indices (SI) for GRE and GMAT Tests

Test	Susceptibility Index: Experimenter	Susceptibility Index: Consensus	Number of Items	Number of Options
GRE-I	.29	.30	15	5
GRE-II	.32	.32	15	5
GRE-III	.43*	.42	14	5
GRE Average	.34**	.35		
GMAT-I	.30*	.31	30	5
GMAT-II	.23	.20	30	5
GMAT Average	.27*	.26		
Overall Average	.30**	.29		

* $p < .05$
** $p < .01$

Table 2

Susceptibility Indices (SI) for Four Mathematics Tests

Test	SI	Number of Items	Number of Options	"Not Given" Type Option?
Iowa Test of Basic Skills				
Test A-1 Arithmetic Concepts	.29	37	4	No
Test A-2 Problem Solving	.42**	34	4	Yes
Average	.35*			
California Achievement Tests				
Section 3 Mathematics Computation	.23	40	5	Yes
Section 4 Mathematics Concepts	.27	25	5	Yes
Section 5 Problems	.28	15	5	Yes
Average	.25			
Differential Aptitude Test				
Numerical Ability	.41**	40	5	Yes
Stanford Diagnostic Mathematics Test				
Test 1 Number System and Numeration	.47***	36	5	No
Test 2 Computation	.25	48	5	Yes
Test 3 Applications	.26	24	5	Yes
Average	.33**			

* $p < .05$
** $p < .01$
*** $p < .001$

mathematics items were applied. All four of these tests have an average SI greater than that expected by chance, and three of these were significant. These results also lend support to Hypothesis 1, indicating that the susceptibility of mathematics tests to test-wiseness is prevalent. When the principles were applied to Test 1 of the Stanford Diagnostic Mathematics Test independently by the consultant, the susceptibility index for that subtest was equal to .46, which is very close to the .47 obtained by the experimenter.

A few clarifications need to be made regarding these tests. For all but one item in Test A-2 of the Iowa Test of Basic Skills, the fourth option was "Not Given" (which was the correct answer for 2 items). For all but 2 of the items on the California Achievement Tests, the fifth option was "None of These" (which was the correct answer for 11 items). The fifth option for all items of the Differential Aptitude Test of Numerical Ability was "None of These" (which was the correct answer for 7 items). Thirty-three of the items on Test 2 and five of the items on Test 3 of the Stanford Diagnostic Mathematics Tests had the fifth option "Not Here" (which was the correct answer for 3 items).

Test 2 of the Stanford Diagnostic Mathematics Test contains two types of items, one type with five numerical options and the other with a fifth option of "Not Here." Contrary to the results reported above, the susceptibility indices (SI) of the items with "Not Here" as the fifth option (SI = .25) were not very different from those items with five numerical options (SI = .26).

It is difficult to compare the susceptibility indices for 4- and

5-option items, since many of the items have an option of or similar to "Not Given;" however, there may be some evidence for lower susceptibility indices among 4-option items, supporting Hypothesis 2. Test A-1 (ITBS) is the only subtest with 4-option items all having four numerical options. The index of .29 for this subtest is not very different from the .25 expected by chance. Sections 3, 4, 5 (CAT), Numerical Ability Test (DAT), and Tests 2 and 3 (SDMT) all had four numerical options plus an option such as "Not Given". If we consider these subtests as having 4-option items, then five of the six subtests have indices around the chance index. (It is not too unreasonable to view these as 4-option items. The frequency with which the "Not Given" or similar option was the correct answer was considerably lower than chance. This may even warrant a "rule" suggesting not to select an option of this type when guessing.) Test 1 (SDMT) was the only subtest with 5-numerical-option items, and the index of .47 is quite high. On Test A-2 (ITBS) consider the index of .42 as an index for three-option items. This index is still above that expected by chance. But, ignoring the "Not Given" option, it is also an index for an odd number of options. One of the unstated premises for suggesting that 4-option items would be less susceptible to test-wiseness than 5-option items was based on the thought that with an even number of options it would be more difficult for a pattern to emerge (e.g., (A) 6 (B) 7 (C) 9 (D) 10 does not offer clues to the correct option as does (A) 6 (B) 7 (C) 9 (D) 10 (E) 12 (see page 7) and (A) 6 (B) 9 (C) 12).

The average SI for all tests examined was greater than that expected by chance. In fact, for all tests, the average SI was as high

or higher than that expected if at least one option were eliminated. The SI for the GRE (SI = .34, $p < .01$) is considered to be moderately high, given that by chance it is expected that only 20 percent of the items should be answered correctly. If this test has a high SI, then, presumably, tests of lower quality would have even higher SI's. Hence, there should be concern about the quality of mathematics items with respect to test-wiseness. There seems to be a definite tendency for item writers to produce mathematics items with patterns among the distractors. In the future, item writers may want to be cognizant of the possibility of producing items susceptible to test-wiseness.

There seems to be some support for the hypothesis that 4-option items are less susceptible to test-wiseness than 5-option items. There was a large difference in the SI's between Test 1 (SDMT), which had all five options numerical (SI = .47, $p < .001$), and Test A-1 (ITBS), which had all 4 options numerical (SI = .29, $p > .05$). The findings here are not strong enough to decide that 4-option items are less susceptible to test-wiseness than 5-option items, but test constructors should keep this possibility in mind.

A Study of Training Test-Wiseness for Mathematics Items

Since the above results confirmed that mathematics items could be susceptible to test-wiseness, the experimental part of this research was carried out. Based on their SAT Math scores, subjects were paired and randomly assigned to one of four groups. Two of the groups, A-1 and A-2, received only the options to items. They were given 80 items in all, 40 items with 5 options and the same 40 items with four options (one option

eliminated). The sets of 4-option and 5-option items were presented in alternating order. The items were randomly selected from disclosed forms of the GRE and GMAT.

One group, A-1, received training in test-wiseness and the other group received no training. These Ss also were asked to complete a brief questionnaire, on which they were asked to provide their reactions to the study.

Since the effectiveness of training in test-wiseness was being compared for 4-option and 5-option items, it was necessary to change the 5-option items to 4-option items. This could not be done by randomly eliminating one of the five options for two reasons. First, the purpose of changing the 5-option items into 4-option items was to break up any patterns leading to the correct answer. Second, for items in which patterns formed would lead the test-wise examinee to choose an incorrect answer, if a misleading pattern was left in an item this could give biased support to hypothesis 5. Thus, it was important to leave each item as free from cues as possible.

The other two groups, B-1 and B-2, of Ss were given the same 40 items, this time with the questions included, and were asked to solve the problems. Group B-1 was given the items with 5 options and Group B-2 was given the 4-option version of the items. Neither group received training in test-wiseness.

Results and Discussion

Means for the total scores on the 5-option items for Groups A-1 and A-2 were compared to the expected mean score (20%) and to each other

by the use of t-tests. The same was done for the 4-option items. These same tests were done separately for the GRE items and the GMAT items. Results appear in Tables 3 and 4.

Table 3

Means and Standard Deviations of Scores for Groups A-1 and A-2

		5-option Items			4-option Items		
Group		GRE	GMAT	Total	GRE	GMAT	Total
A-1 (n=21) Trained	$\bar{X}$=	5.62	3.81	9.43	4.52	4.86	9.38
	s=	1.66	1.40	1.96	1.40	2.22	2.36
A-2 (n=21) Untrained	$\bar{X}$=	3.86	4.43	8.29	4.76	4.91	9.67
	s=	1.53	1.72	1.71	1.76	2.17	2.90
Number of Items		20	20	40	20	20	40
Score Expected by Chance		4	4	8	5	5	10
Score Expected if Rules Followed		6.1	5.4	11.5	*	*	*

*Cues eliminated in the 4-option version

As was predicted, Ss receiving no training in test-wiseness did not score significantly different than expected by chance. It was predicted that Ss trained in test-wiseness would score significantly higher than chance on 5-option items but not on 4-option items. For the 5-option items, Ss trained in test-wiseness did score significantly higher than chance on the total test and the GRE subset of items, but not on the GMAT subset of items. Since Table 1 indicates that the GMAT test is not cue-infested, or highly susceptible to test-wiseness,

it is not surprising that the GMAT results are not significant. On the 4-option items, the trained Ss did not score significantly higher than expected by chance. It was also hypothesized that the trained Ss would score higher than the untrained Ss on the 5-option items but not on the 4-option items. For 4-option items this was the case. For 5-option items, the trained Ss' total scores were not significantly higher than those for the untrained Ss, but the scores on the GRE subset of items were significantly higher for the trained Ss.

Table 4

t Values for Comparing Groups A-1 and A-2

	5-option Items			4-option Items		
Comparison	GRE	GMAT	Total	GRE	GMAT	Total
A-1 vs Chance	t= 4.36*	-.61	3.26**	-1.53	-.28	-1.17
A-2 vs Chance	t= -.41	1.12	.76	-.61	-.21	-.51
A-1 vs A-2	t= 3.31*	1.40	1.79	-.48	-.06	-.35
	n= 20	20	40	20	20	40

* $p < .01$
** $p < .001$

On the self-administered questionnaire, only one S in Group A-1 circled "No" to the first question asking if the rules for guessing the correct option were understandable. Later this S clarified that the rules were understandable, but the purpose of the study was not clear, and that is why "No" was circled. The rules used most often by this group were those involving the middle of a sequence, the reciprocal,

common elements, and "odd" options. Half of this group (voluntarily) stated that it was more difficult to apply rules to 4-option items than to 5-option items.

Over one-half of the Ss in Group A-2 reported that they used specific reasoning strategies to select the correct answer. These reported strategies coincided with many of the rules presented to Group A-1. For example, one S in Group A-2 wrote:

> I tended to not circle the extreme (highest or lowest) answers. I also tended not to circle answers that were very different from the others--powers of ten higher than all the rest, not containing fractions when all the others did . . . . I did circle those that appeared to be multiples of the others, and those in ~~the~~ middle of the range.

Based on the procedures put forth by Ebel (1965, p. 347), indices of discrimination for each item for Groups B-1 and B-2 were calculated. Then, for each item, this index for the 5-option items was compared to the index for the corresponding 4-option items, using a t-test for related samples. Results appear in Tables 5 and 6. In addition, reliability indices were computed using Kuder-Richardson formula 20 for the two tests, and these indices appear in Table 7.

There were no significant differences between the indices of discrimination of the 4- and 5-option items, as was predicted. This was true for the total test, as well as the GRE and GMAT subtests. The reliabilities for the total tests were not significantly different for the 4- and 5-option tests.

Table 5

Means and Standard Deviations of Scores
for Groups B-1 and B-2

Group		GRE	GMAT	Total
B-1 (n=21)	$\bar{X}$=	15.38	10.48	25.86
5-option	s=	3.25	3.64	6.41
B-2 (n=21)	$\bar{X}$=	15.10	11.43	26.53
4-option	s=	2.19	4.21	5.91
Number of Items		20	20	40

Table 6

Means*, Standard Deviations*, and t Values
Comparing Indices of Discrimination for
4-option Items and 5-option Items

	4-option Test			5-option Test		
	GRE	GMAT	Total	GRE	GMAT	Total
$\bar{X}$=	.40	.44	.41	.27	.67	.49
s=	.25	.30	.27	.36	.72	.53

Comparison between 4- and 5-option Tests

Test	t	df
GRE	1.57	19
GMAT	-1.01	19
Total	1.51	39

* Note that an r to Z transformation was used to compute the means and standard deviations of the discrimination indices.

Table 7

Reliability Indices of Tests Given
Groups B-1 and B-2

	GRE	GMAT	Total
5-option Test	.76	.75	.83
4-option Test	.44	.84	.84

Conclusions

The results of the studies presented in this work showed some support for all seven of the hypotheses this work set out to test. The investigator finds these results encouraging in that her hypotheses were supported, and at the same time, discouraging because of the possible implications that can be drawn.

The susceptibility indices for the tests analyzed indicate that mathematics items are flawed in that they are susceptible to test-wiseness. As Ebel (1972) notes, flawed items introduce additional variance into tests, and this additional variance is not desirable. In general, two methods for dealing with test-wiseness are suggested: 1) training all potential examinees in test-wiseness and 2) eliminating items susceptible to test-wiseness (Sarnacki, 1979). Since this investigator supports eliminating test-wiseness cues, the former suggestion will be discussed only briefly.

This study found that examinees could be trained in test-wiseness for mathematics items. The principles of test-wiseness developed for this study based on the GRE and GMAT tests were readily applied to other

tests, and were easily understood by the Ss receiving training in test-wiseness. These principles could be used as a basis for training examinees in test-wiseness in the future. These principles should be refined, however, based on the frequency with which each principle is used and on strategies used by the Ss not trained in test-wiseness.

One way to start eliminating test-wiseness is for item writers to be cognizant of the principles put forth in this work when constructing distractors. Another way to start eliminating items susceptible to test-wiseness may be to reduce the number of options. This study also showed that flaws in the items could be eliminated by eliminating an option based on these principles. The Ss trained in test-wiseness did not score better than chance on these 4-option items. Results of this study and other studies have also shown that various psychometric properties are not significantly affected by reducing the number of options. If item writers are required to write fewer distractors, more effort can be paid to each distractor. Also, if there are fewer distractors, then it is likely that fewer principles of test-wiseness would need to be checked against the distractors for possible patterns. Besides wanting information on the effect on other test statistics, the test constructor will want to know the best way to determine which four options should be put into an item.

In conclusion, although many existing principles of test-wiseness are not apparently applicable to mathematics items, a modified interpretation of some of these principles can be applied to items with numerical options. These modified principles are not being followed

in well-respected standardized tests currently being used and may make it possible for some examinees to invalidly inflate their scores. Evidence from this research suggests that the psychometric quality of the test is not materially affected by correcting flawed items and reducing the number of options from five to four.

References

Ebel, Robert L. Measuring Educational Achievement. Englewood Cliffs, N.J.: Prentice-Hall, 1965.

Ebel, Robert L. Essentials of Educational Measurement (second edition). Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Miller, Harry G., Williams, Reed G., and Haladyna, Thomas M. Beyond Facts: Objective Ways to Measure Thinking. Englewood Cliffs, N.J.: Educational Technology Publications, 1978.

Millman, Jason, Bishop, Carol H., and Ebel, Robert, "An Analysis of Test-Wiseness," Educational and Psychological Measurement, 1965, 25, 707-726.

Millman, Jason and Pauk, Walter. How to Take Tests. New York: McGraw Hill, Inc., 1969.

Mosier, Charles I., Myers, M. Claire, and Price, Helen G., "Suggestions for the Construction of Multiple-Choice Items," in Chase, Clinton I., and Ludlow, H. Glenn, Ed., Readings in Educational and Psychological Measurement. Boston: Houghton Mifflin, Co., 1966.

Ramos, Robert A., and Stern, June, "Item Behavior Associated With Changes in the Number of Alternatives in Multiple Choice Items," Journal of Educational Measurement, 1973, 10, 305-310.

Sarnacki, Randolph E., "An Examination of Test-Wiseness in the Cognitive Test Domain," Review of Educational Research, 1979, 49, 252-279.

Smith, Jeffrey K., Proposal for "A Cognitive Model of Test Taking Skills," Rutgers University, October 1980.

Tversky, A., "On the Optimal Number of Alternatives of a Choice Point," Journal of Mathematical Psychology, 1964, 1, 386-391.